

**МИНИСТЕРСТВО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ**

**КАЗАНСКИЙ ГОСУДАРСТВЕННЫЙ  
ФИНАНСОВО-ЭКОНОМИЧЕСКИЙ ИНСТИТУТ**

**Кафедра прогнозирования и статистики**

## **КОНСПЕКТ ЛЕКЦИЙ**

**по курсу «ЭКОНОМЕТРИКА»  
для студентов III курса дневного отделения  
всех специальностей**

**Часть I**

**Казань 2003**

Печатается по решению кафедры прогнозирования и статистики: **протокол №5 от 20.01.03.**

**Автор: доцент Костромин Андрей Владиленович**

## Введение

**Эконометрика** – это наука, в которой на базе реальных статистических данных строятся, анализируются и совершенствуются математические модели реальных экономических явлений. Эконометрика позволяет найти количественное подтверждение либо опровержение того или иного экономического закона либо гипотезы.

Эконометрика как научная дисциплина зародилась и получила развитие на основе слияния экономической теории, математической экономики и экономической и математической статистики.

По словам Р. Фриша: «... каждая из трех отправных точек - статистика, экономическая теория и математика-необходимое, но не достаточное условие для понимания количественных соотношений в современной экономической жизни. Это - единство всех трех составляющих. И это единство образует эконометрику».

Таким образом, **эконометрика - это наука, которая дает количественное выражение взаимосвязей экономических явлений и процессов.**

**Предметом** эконометрики являются экономические явления. Однако, в отличие от экономической теории, эконометрика делает упор на количественные, а не на качественные аспекты этих явлений. Например, известно, что спрос на товар с ростом его цены падает. Однако, как быстро и по какому закону это происходит, в экономической теории не определяется. Это в каждом конкретном случае делает эконометрика. С другой стороны, математическая экономика строит и анализирует модели экономических процессов без использования реальных числовых значений. Эконометрика же изучает модели на базе эмпирических данных.

Наконец, в эконометрике широко используется аппарат математической статистики, особенно при установлении связей между экономическими показателями. В то же время в экономике невозможно проведение управляемого эксперимента, и эконометристы используют свои собственные приемы анализа, которые в математической статистике не встречаются.

Основными **целями** эконометрики являются:

**1. Прогноз** экономических и социально-экономических показателей, характеризующих состояние и развитие анализируемой системы.

**2. Имитация** различных возможных сценариев социально-экономического развития.

### **Основные задачи эконометрики:**

1. Построение эконометрических моделей, т.е. представление экономических моделей в математической форме, удобной для проведения эмпирического анализа (**спецификация модели**).

2. Оценка параметров построенной модели, делающих выбранную модель наиболее адекватной реальным данным (**параметризация**).

3. Проверка качества найденных параметров модели и самой модели в целом (**верификация**).

4. Использование построенных моделей для объяснения поведения экономических показателей, прогнозирования и предсказания, а также для осмысленного проведения экономической политики.

### **Этапы эконометрического моделирования:**

1. **Постановочный этап**: определение конечных целей моделирования, набора факторов и показателей.

2. **Априорный этап**: предмодельный анализ экономической сущности изучаемого явления.

3. **Параметризация**: собственно моделирование, т.е. выбор общего вида модели, состава и формы входящих в нее связей.

4. **Информационный этап**: сбор статистической информации.

5. **Идентификация модели**: статистический анализ модели и оценивание неизвестных параметров модели.

6. **Верификация модели**: сопоставление реальных и модельных данных, проверка адекватности модели, оценка точности модельных данных.

## **ПАРНАЯ РЕГРЕССИЯ**

### **Спецификация модели**

В зависимости от количества факторов, включенных в уравнение регрессии, принято различать простую (парную) и множественную регрессию.

**Простая регрессия** представляет собой регрессию между двумя переменными –  $y$  и  $x$ , т.е. модель вида:

$$y = \hat{f}(x), \quad (1)$$

где  $y$ -зависимая переменная (результативный признак);  $x$ - независимая, или объясняющая переменная (признак – фактор, или регрессор).

**Множественная регрессия** представляет собой регрессию результативного признака с двумя и большим числом факторов, т.е. модель вида:

$$y = \hat{f}(x_1, x_2, \dots, x_k) \quad (2)$$

Любое эконометрическое исследование начинается со **спецификации модели**, т.е. с формулировки вида модели, исходя из соответствующей теории связи между переменными.

Из всего круга факторов, влияющих на результативный признак, необходимо выделить наиболее существенно влияющие факторы. Парная регрессия достаточна, если имеется доминирующий фактор, который и используется в качестве объясняющей переменной. Например, выдвигается гипотеза о том, что величина спроса  $y$  на товар находится в обратной зависимости от цены  $x$ , т.е.  $\hat{y}_x = a - b \cdot x$ .

Уравнение простой регрессии характеризует связь между двумя переменными, которая проявляется как закономерность лишь **в среднем** по совокупности наблюдений. (Например, если зависимость спроса  $y$  от цены  $x$  имеет вид:  $y = 5000 - 2 \cdot x$  Это означает, что с ростом цены на 1 д.е. спрос **в среднем** уменьшается на 2 д.е.).

В уравнении регрессии корреляционная по сути связь признаков представляется в виде функциональной связи. В каждом отдельном случае величина  $y$  складывается из двух слагаемых:

$$y_j = \hat{y}_{x_j} + \varepsilon_j,$$

где  $y_j$  – фактическое значение результативного признака;  $\hat{y}_{x_j}$  – значение признака, найденное из математической функции связи  $y$  и  $x$ , т.е. из уравнения регрессии;  $\varepsilon_j$  – случайная величина, характеризующая отклонение реального значения признака от найденного по уравнению регрессии.

Случайная величина  $\varepsilon$  называется также **возмущением**. Она включает влияние не учтенных в модели факторов, случайных ошибок и особенностей измерения. Ее порождают 3 источника: спецификация модели, выборочный характер исходных данных и ошибки измерения.

Например, зависимость спроса от цены точнее следует записывать так:

$$y = 5000 - 2 \cdot x + \varepsilon$$

В данном случае слева записано просто  $y$ , что означает фактическое значение, а не  $\hat{y}$ , отвечающее значению, рассчитанному по уравнению регрессии.

**Ошибки спецификации.** Это прежде всего неправильно выбранная форма модели. В частности, зависимость спроса от цены может быть выражена линейно

$$\hat{y}_x = a - b \cdot x,$$

но возможны и другие соотношения, например

$$\hat{y} = ax^{-b}, \quad \hat{y} = a + \frac{b}{x}, \quad \hat{y} = \frac{1}{a + bx}.$$

Ошибки спецификации тем меньше, чем в большей мере теоретические значения признака подходят к фактическим данным  $y$ .

К ошибкам спецификации относится также недоучет в уравнении регрессии какого-либо существенного фактора, т.е. использование парной регрессии вместо множественной. Например, спрос на конкретный товар может определяться не только ценой, но и доходом на душу населения.

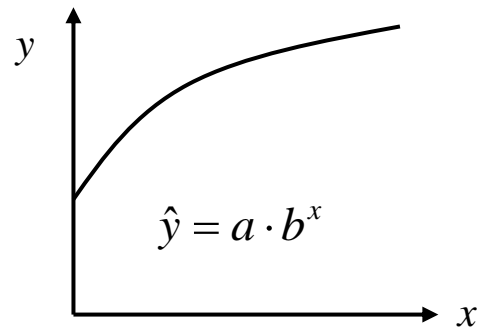
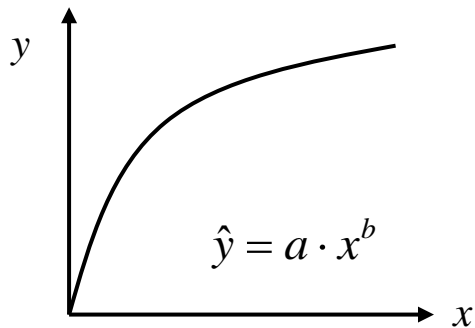
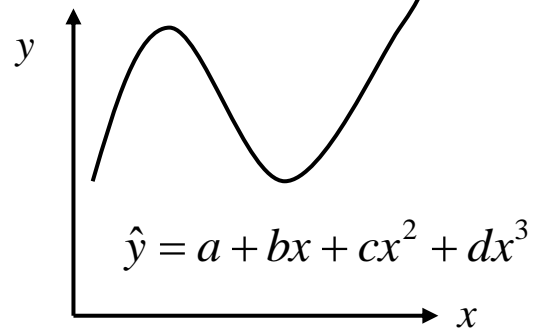
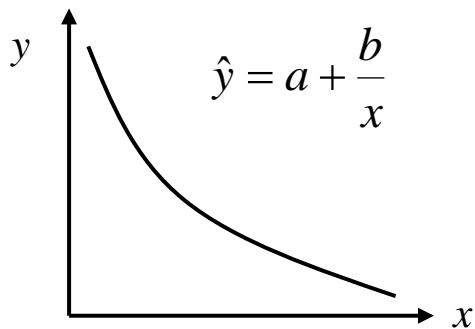
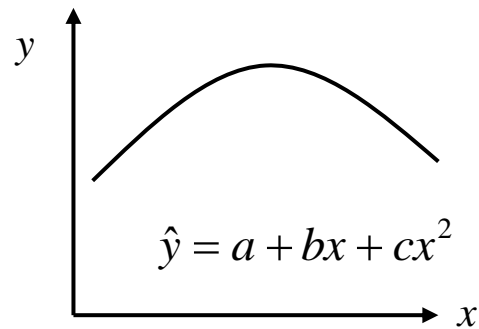
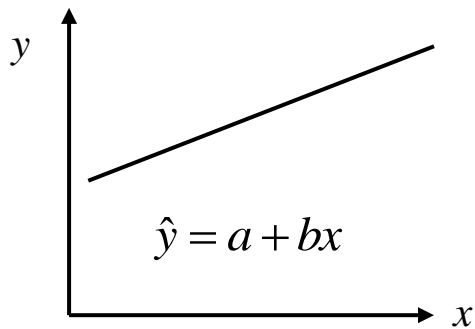
**Ошибки выборки.** Исследователь при установлении связи между признаками имеет дело с выборочными данными. При изучении экономических процессов данные в исходной совокупности часто являются неоднородными. В этом случае уравнение регрессии не имеет практического смысла. Поэтому для получения хорошего результата из выборки исключают данные с аномальными значениями исследуемых признаков.

**Ошибки измерения.** Представляют наибольшую опасность в практическом использовании методов регрессии. Ошибки спецификации можно уменьшить, изменяя форму модели, ошибки выборки - увеличивая объем исходных данных, ошибки измерения сводят на нет все усилия по количественной оценке связи между признаками. Например, статистическое измерение дохода на душу населения может иметь ошибку в результате наличия сокрытых доходов. Другой пример: органы государственной статистики получают балансы предприятий, достоверность которых никто не подтверждает.

В эконометрических исследованиях предполагается, что ошибки измерения сведены к минимуму. Поэтому основное внимание уделяется ошибкам спецификации модели.

В парной регрессии выбор вида математической функции (1) может быть осуществлен тремя методами: графическим, аналитическим и экспериментальным.

**Графический метод** достаточно нагляден. Он основан на поле корреляции. Рассмотрим типы кривых.



Используются и другие типы кривых:

$$\hat{y} = \frac{1}{a + b \cdot x}; \quad \hat{y} = a + bx + \frac{c}{x}; \quad \hat{y} = a + b \cdot \lg x;$$

$$\hat{y} = \frac{1}{a + bx + cx^2}; \quad \hat{y} = \frac{a}{1 + be^{-cx}}; \quad \lg \hat{y} = a + bx + cx^2.$$

**Аналитический метод** выбора типа уравнения регрессии основан на изучении материальной природы связи исследуемых признаков.

Пусть, например, изучается потребность предприятия в электроэнергии  $y$  в зависимости от объема выпускаемой продукции  $x$ . Все потребление электроэнергии можно подразделить на 2 части:

- не связанное с производством продукции  $a$ ;
- непосредственно связанное с объемом выпускаемой продукции, пропорционально возрастающее с увеличением объема выпуска  $bx$ ;

Тогда зависимость потребления электроэнергии от объема продукции можно выразить уравнением регрессии вида

$$\hat{y} = a + bx$$

Разделив на  $x$ , получим удельный расход электроэнергии на единицу продукции  $z_x = y/x$ :

$$\hat{z}_x = b + \frac{a}{x}$$

Это равносторонняя гипербола.

Аналогично затраты предприятия могут быть условно-переменные, изменяющиеся пропорционально изменению объема продукции (расход материала, оплата труда и др.) и условно-постоянные, не изменяющиеся с изменением объема производства (арендная плата, содержание администрации и др.). Соответствующая зависимость затрат на производство  $y$  от объема продукции  $x$  характеризуется линейной функцией.

$$y = a + bx,$$

а зависимость себестоимости единицы продукции  $z_x$  от объема продукции – равносторонней гиперболой:

$$\hat{z}_x = b + \frac{a}{x}$$

**Экспериментальный метод** используется при обработке информации на компьютере путем сравнения величины остаточной дисперсии  $D_{ост}$ , рассчитанной на разных моделях. В практических исследованиях, как правило, имеет место некоторое рассеяние точек относительно линии регрессии. Оно обусловлено влиянием прочих, не учитываемых в уравнении регрессии факторов:

$$D_{ост} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_{x_i})^2$$

Чем меньше  $D_{ост}$ , тем меньше наблюдается влияние прочих факторов, тем лучше уравнение регрессии подходит к исходным данным. При обработке данных на компьютере разные математические



функции перебираются в автоматическом режиме, и из них выбирается та, для которой  $D_{ост}$  является наименьшей.

Если  $D_{ост}$  примерно одинакова для нескольких функций, то на практике выбирают более простую, так как она в большей степени поддается интерпретации и требует меньшего объема наблюдений. Результаты многих исследований подтверждают, что число наблюдений должно в 6-7 раз превышать число рассчитываемых параметров при переменной  $x$ . Это означает, что искать линейную регрессию, имея менее 7 наблюдений, вообще не имеет смысла. Если вид функции усложняется, то требуется увеличение объема наблюдений. Для рядов динамики, ограниченных по протяженности – 10, 20, 30 лет – предпочтительна модель с меньшим числом параметров при  $x$ .

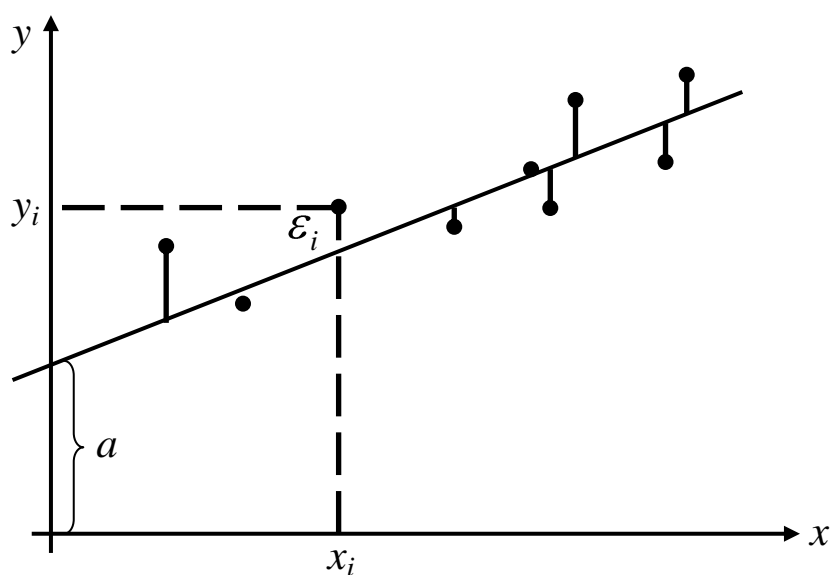
### Оценка параметров линейной регрессии.

Линейная регрессия сводится к нахождению уравнения вида

$$\hat{y}_x = a + bx \quad (\text{или } y = a + bx + \varepsilon) \quad (3)$$

Первое выражение позволяет по заданным значениям фактора  $x$  рассчитать теоретические значения результативного признака, подставляя в него фактические значения фактора  $x$ . На графике теоретические значения лежат на прямой, которая представляет собой линию регрессии.

Построение линейной регрессии сводится к оценке ее параметров  $a$  и  $b$ . Классический подход к оцениванию параметров линейной регрессии основан на **методе наименьших квадратов (МНК)**.



МНК позволяет получить такие оценки параметров  $a$  и  $b$ , при которых сумма квадратов отклонений фактических значений  $y$  от теоретических  $\hat{y}_x$  минимальна:

$$\sum_i (y_i - \hat{y}_{x_i})^2 \rightarrow \min, \text{ или } \sum_{i=1}^n \varepsilon_i^2 \rightarrow \min \quad (4)$$

Для нахождения минимума надо вычислить частные производные суммы (4) по каждому из параметров -  $a$  и  $b$  - и приравнять их к нулю.

$$\Sigma \varepsilon_i^2 = S^2 = \Sigma (y - \hat{y}_x)^2 = \Sigma (y - a - bx)^2$$

$$\begin{cases} \frac{\partial S}{\partial a} = -2 \Sigma y + 2na + 2b \Sigma x = 0; \\ \frac{\partial S}{\partial b} = -2 \Sigma y \cdot x + 2a \Sigma x + 2b \Sigma x^2 = 0 \end{cases} \quad (5)$$

Преобразуем, получаем **систему нормальных уравнений**:

$$\begin{cases} n \cdot a + b \Sigma x = \Sigma y, \\ a \Sigma x + b \Sigma x^2 = \Sigma yx \end{cases} \quad (6)$$

В этой системе  $n$ - объем выборки, суммы легко рассчитываются из исходных данных. Решаем систему относительно  $a$  и  $b$ , получаем:

$$b = \frac{n \Sigma yx - \Sigma y \Sigma x}{n \Sigma x^2 - (\Sigma x)^2}, \quad (7)$$

$$a = \frac{1}{n} \Sigma y - \frac{b}{n} \Sigma x. \quad (8)$$

Выражение (7) можно записать в другом виде:

$$b = \frac{\overline{yx} - \bar{y} \cdot \bar{x}}{\overline{x^2} - \bar{x}^2} = \frac{cov(x, y)}{\sigma_x^2}, \quad (9)$$

где  $cov(x, y)$  — ковариация признаков,  $\sigma_x^2$  — дисперсия фактора  $x$ .

Параметр  $b$  называется **коэффициентом регрессии**. Его величина показывает среднее изменение результата с изменением фактора на одну единицу. Возможность четкой экономической интерпретации

коэффициента регрессии сделала линейное уравнение регрессии достаточно распространенным в эконометрических исследованиях.

Формально  $a$ - значение  $y$  при  $x=0$ . Если  $x$  не имеет и не может иметь нулевого значения, то такая трактовка свободного члена  $a$  не имеет смысла. Параметр  $a$  может не иметь экономического содержания. Попытки экономически интерпретировать его могут привести к абсурду, особенно при  $a < 0$ . Интерпретировать можно лишь знак при параметре  $a$ . Если  $a > 0$ , то относительное изменение результата происходит медленнее, чем изменение фактора. Сравним эти относительные изменения:

$$bx < a + bx \text{ при } a > 0, x > 0 \Rightarrow b < \frac{a + bx}{x}$$

$$\Rightarrow \frac{bdx}{dx} < \frac{a + bx}{x} \Rightarrow \frac{dy}{dx} < \frac{y}{x} \Rightarrow \frac{dy}{y} < \frac{dx}{x}.$$

Иногда линейное уравнение парной регрессии записывают для отклонений от средних значений:

$$y' = b \cdot x', \quad (10)$$

где  $y' = y - \bar{y}$ ,  $x' = x - \bar{x}$ . При этом свободный член равен нулю, что и отражено в выражении (10). Этот факт следует из геометрических соображений: уравнению регрессии отвечает та же прямая (3), но при оценке регрессии в отклонениях начало координат перемещается в точку с координатами  $(\bar{x}, \bar{y})$ . При этом в выражении (8) обе суммы будут равны нулю, что и повлечет равенство нулю свободного члена.

Рассмотрим в качестве примера по группе предприятий, выпускающих один вид продукции, функцию издержек  $y = a + bx + \varepsilon$

Табл. 1.

Выпуск продукции тыс.ед.( $x$ )	Затраты на производство, млн.руб.( $y$ )	$y \cdot x$	$x^2$	$y^2$	$\hat{y}_x$
1	30	30	1	900	31,1
2	70	140	4	4900	67,9
4	150	600	16	22500	141,6
3	100	300	9	10000	104,7
5	170	850	25	28900	178,4
3	100	300	9	10000	104,7
4	150	600	16	22500	141,6
Итого: 22	770	2820	80	99700	770,0

Система нормальных уравнений будет иметь вид:

$$\begin{cases} 7a + 22b = 770 \\ 22a + 80b = 2820 \end{cases}$$

Решая её, получаем  $a = -5,79$ ,  $b = 36,84$ .

Уравнение регрессии имеет вид:

$$\hat{y}_x = -5,79 + 36,84x$$

Подставив в уравнение значения  $x$ , найдем теоретические значения  $y$  (последняя колонка таблицы).

Величина  $a$  не имеет экономического смысла. Если переменные  $x$  и  $y$  выразить через отклонения от средних уровней, то линия регрессии на графике пройдет через начало координат. Оценка коэффициента регрессии при этом не изменится:

$$\hat{y}' = 36,84 \cdot x', \text{ где } y' = y - \bar{y}, x' = x - \bar{x}.$$

В качестве другого примера рассмотрим функцию потребления в виде:

$$C = K \cdot y + L,$$

где  $C$  - потребление,  $y$  - доход,  $K, L$  - параметры. Данное уравнение линейной регрессии обычно используется в увязке с балансовым равенством:

$$y = C + I - r,$$

где  $I$  - размер инвестиций,  $r$  - сбережения.

Для простоты предположим, что доход расходуется на потребление и инвестиции. Таким образом, рассматривается система уравнений:

$$\begin{cases} C = K \cdot y + L, \\ y = C + L. \end{cases}$$

Наличие балансового равенства накладывает ограничения на величину коэффициента регрессии, которая не может быть больше единицы, т.е.  $K \leq 1$ .

Предположим, что функция потребления составила:

$$\hat{C} = 1,9 + 0,65 \cdot y.$$

Коэффициент регрессии характеризует склонность к потреблению. Он показывает, что из каждой тысячи рублей дохода на потребление расходуется в среднем 650 руб., а 350 руб. инвестируется. Если рассчитать регрессию размера инвестиций от дохода, т.е.  $\hat{I} = a + by$ , то уравнение регрессии составит  $\hat{I} = 1,9 + 0,35y$ . Это уравнение можно и не определять, поскольку оно выводится из функции потреб-

ления. Коэффициенты регрессии этих двух уравнений связаны равенством:

$$0,65 + 0,35 = 1.$$

Если коэффициент регрессии оказывается больше единицы, то  $y < C + I$ , и на потребление расходуются не только доходы, но и сбережения.

Коэффициент регрессии в функции потребления используется для расчета мультипликатора:

$$m = \frac{1}{1 - b}.$$

Здесь  $m \approx 2,86$ , поэтому дополнительные вложения 1 тыс. руб. на длительный срок приведут при прочих равных условиях к дополнительному доходу 2,86 тыс. руб.

При линейной регрессии в качестве показателя тесноты связи выступает линейный коэффициент корреляции  $r$ :

$$r = b \frac{\sigma_x}{\sigma_y} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{\overline{yx} - \bar{y} \cdot \bar{x}}{\sigma_x \cdot \sigma_y}$$

Его значения находятся в границах:  $-1 < r \leq 1$ . Если  $b > 0$ , то  $0 \leq r \leq 1$ , при  $b < 0$   $-1 \leq r < 0$ . По данным примера  $r = 0,991$ , что означает очень тесную зависимость затрат на производство от величины объема выпускаемой продукции.

Для оценки качества подбора линейной функции рассчитывается **коэффициент детерминации** как квадрат линейного коэффициента корреляции  $r^2$ . Он характеризует долю дисперсии результативного признака  $y$ , объясняемую регрессией, в общей дисперсии результативного признака:

$$r^2 = \frac{\sigma_{\text{объясн.}}^2}{\sigma_{y \text{ общ.}}^2}$$

Величина  $1 - r^2$  характеризует долю дисперсии  $y$ , вызванную влиянием остальных, не учтенных в модели факторов.

В примере  $\sigma^2 = 0,092$ . Уравнением регрессии объясняется 98,2 % дисперсии  $y$ , а на прочие факторы приходится 1,8 %, это остаточная дисперсия.

### Предпосылки МНК (условия Гаусса-Маркова)

Как было сказано выше, связь между  $y$  и  $x$  в парной регрессии является не функциональной, а корреляционной. Поэтому оценки параметров  $a$  и  $b$  являются случайными величинами, свойства которых существенно зависят от свойств случайной составляющей  $\varepsilon$ . Для получения по МНК наилучших результатов необходимо выполнение следующих предпосылок относительно случайного отклонения (условия Гаусса – Маркова):

1<sup>0</sup>. Математическое ожидание случайного отклонения равно нулю для всех наблюдений:  $M \varepsilon_i = 0, \forall i$ .

2<sup>0</sup>. Дисперсия случайных отклонений постоянна:  $D \varepsilon_i = D \varepsilon_j = \sigma^2, \forall i, j$ .

Выполнимость данной предпосылки называется **гомоскедастичностью** (постоянством дисперсии отклонений). Невыполнимость данной предпосылки называется **гетероскедастичностью** (непостоянством дисперсии отклонений)

3<sup>0</sup>. Случайные отклонения  $\varepsilon_i$  и  $\varepsilon_j$  являются независимыми друг от друга для  $i \neq j$ :

$$\text{cov} \varepsilon_i, \varepsilon_j = \begin{cases} 0, & i \neq j \\ \sigma^2, & i = j. \end{cases}$$

Выполнимость этого условия называется **отсутствием автокорреляции**.

4<sup>0</sup>. Случайное отклонение должно быть независимо от объясняющих переменных.

Обычно это условие выполняется автоматически, если объясняющие переменные в данной модели не являются случайными. Кроме того, выполнимость данной предпосылки для эконометрических моделей не столь критична по сравнению с первыми тремя.

При выполнении указанных предпосылок имеет место **теорема Гаусса-Маркова: оценки (7) и (8), полученные по МНК, имеют наименьшую дисперсию в классе всех линейных несмещенных оценок**.

Таким образом, при выполнении условий Гаусса-Маркова оценки (7) и (8) являются не только несмещенными оценками коэффициентов регрессии, но и наиболее эффективными, т.е. имеют наименьшую дисперсию по сравнению с любыми другими оценками данных параметров, линейными относительно величин  $y_i$ .

Именно понимание важности условий Гаусса-Маркова отличает компетентного исследователя, использующего регрессионный анализ, от некомпетентного. Если эти условия не выполнены, исследователь должен это сознавать. Если корректирующие действия возможны, то аналитик должен быть в состоянии их выполнить. Если ситуацию исправить невозможно, исследователь должен быть способен оценить, насколько серьезно это может повлиять на результаты.

### **Оценка существенности параметров линейной регрессии и корреляции**

После того, как найдено уравнение линейной регрессии (3), проводится оценка значимости как уравнения в целом, так и отдельных его параметров.

Оценка значимости уравнения регрессии в целом дается с помощью F- критерия Фишера. При этом выдвигается нулевая гипотеза  $H_0: b = 0$  о том, что коэффициент регрессии равен нулю и, следовательно, фактор  $x$  не оказывает влияния на результат  $y$ .

Перед расчетом критерия проводятся анализ дисперсии. Можно показать, что общая сумма квадратов отклонений (СКО)  $y$  от среднего значения  $\bar{y}$  раскладывается на две части – объясненную и необъясненную:

$$\sum (y - \bar{y})^2 = \sum (y - \bar{y})^2 + \sum (y - \hat{y})^2 \quad (11)$$

или, соответственно:

$$(\text{Общая СКО}) = \left( \begin{array}{c} \text{Объясненная} \\ \text{(факторная) СКО} \end{array} \right) + \left( \begin{array}{c} \text{Необъясненная} \\ \text{(остаточная) СКО} \end{array} \right)$$

Здесь возможны два крайних случая: когда общая СКО в точности равна остаточной и когда общая СКО равна факторной.

В первом случае фактор  $x$  не оказывает влияния на результат, вся дисперсия  $y$  обусловлена в основном воздействием прочих факторов, линия регрессии параллельна оси  $Ox$  и  $\bar{y} = \hat{y}$ .

Во втором случае прочие факторы не влияют на результат,  $y$  связан с  $x$  функционально, и остаточная СКО равна нулю.

Однако на практике в правой части (11) присутствуют оба слагаемых. Пригодность линии регрессии для прогноза зависит от того, какая часть общей вариации  $y$  приходится на объясненную вариацию. Если объясненная СКО будет больше остаточной СКО, то уравнение

регрессии статистически значимо и фактор  $x$  оказывает существенное воздействие на результат  $y$ . Это равносильно тому, что коэффициент детерминации будет приближаться к единице.

**Число степеней свободы.** (df-degrees of freedom)- это число независимо варьируемых значений признака.

Для общей СКО требуется  $(n-1)$  независимых отклонений, т.к.  $\sum (y - \bar{y}) = 0$ , что позволяет свободно варьировать  $(n-1)$  значений, а последнее  $n$ -е отклонение определяется из общей суммы, равной нулю. Поэтому  $df_{общ.} = n - 1$ .

Факторную СКО можно выразить так:

$$\begin{aligned} \sum (x - \bar{y})^2 &= \sum (k + bx - (k + b\bar{x}))^2 \\ &= \sum (x - b\bar{x})^2 = b^2 \sum (x - \bar{x})^2 \end{aligned}$$

Эта СКО зависит только от одного параметра -  $b$ , поскольку выражение под знаком суммы к значениям результативного признака не относится. Следовательно, факторная СКО имеет одну степень свободы, и  $df_{факт.} = 1$ .

Для определения  $df_{остат.}$  воспользуемся аналогией с балансовым равенством (11). Так же, как и в равенстве (11), можно записать равенство и между числами степеней свободы:

$$df_{общ.} = df_{факт.} + df_{остат.} \quad (12)$$

Таким образом, можем записать:

$$n - 1 = 1 + n - 2$$

Из этого баланса определяем, что  $df_{остат.} = n - 2$ .

Разделив каждую СКО на свое число степеней свободы, получим **средний квадрат отклонений**, или **дисперсию на одну степень свободы**:

$$\begin{aligned} D_{общ.} &= \frac{\sum (y - \bar{y})^2}{n - 1} \\ D_{факт.} &= \frac{\sum (x - \bar{y})^2}{1} \\ D_{ост.} &= \frac{\sum (y - \hat{y}_x)^2}{n - 2} \end{aligned}$$

Сопоставляя факторную и остаточную дисперсии в расчете на одну степень свободы, получим  $F$  - критерий для проверки нулевой



гипотезы, которая в данном случае записывается как  $H_0 : D_{\text{факт.}} = D_{\text{ост.}}$ .

$$F = \frac{D_{\text{факт.}}}{D_{\text{ост.}}} \quad (13)$$

Если  $H_0$  справедлива, то дисперсии не отличаются друг от друга. Для  $H_0$  необходимо опровержение, чтобы факторная дисперсия превышала остаточную в несколько раз. Английским статистиком Снедекором разработаны таблицы критических значений  $F$  при разных уровнях существенности  $H_0$  и различных числах степеней свободы. Табличное значение  $F$ -критерия – это максимальная величина отношения дисперсий, которая может иметь место при случайном их расхождении для данного уровня вероятности наличия нулевой гипотезы. При нахождении табличного значения  $F$ -критерия задается уровень значимости (обычно 0,05 или 0,01) и две степени свободы – числителя (она равна единице) и знаменателя, равная  $n-2$ .

Вычисленное значение  $F$  признается достоверным (отличным от единицы), если оно больше табличного т.е.  $F_{\text{факт.}} > F_{\text{табл.}}(\alpha; 1; n-2)$ . В этом случае  $H_0$  отклоняется и делается вывод о существенности превышения  $D_{\text{факт.}}$  над  $D_{\text{остат.}}$ , т.е. о существенности статистической связи между  $y$  и  $x$ .

Если  $F_{\text{факт.}} < F_{\text{табл.}}$ , то вероятность  $H_0$  выше заданного уровня (например, 0,05), и эта гипотеза не может быть отклонена без серьезного риска сделать неправильный вывод о наличии связи между  $y$  и  $x$ . Уравнение регрессии считается статистически незначимым,  $H_0$  не отклоняется.

В рассмотренном примере:

$$\sum (y - \bar{y})^2 = \sum y^2 - n \cdot \bar{y}^2 = 99700 - 7 \cdot 110^2 = 15000 - \text{это общая СКО.}$$

$$\sum (x - \bar{x})^2 = b^2 \sum (y - \bar{y})^2 = 36,84^2 \cdot \left( 80 - 7 \cdot \left( \frac{22}{7} \right)^2 \right) = 14735$$

- это факторная СКО.

$$\sum (y - \hat{y}_x)^2 = 15000 - 14735 = 265 - \text{остаточная СКО.}$$

$$D_{\text{факт.}} = 14735; D_{\text{ост.}} = \frac{265}{5} = 53; F_{\text{факт.}} = \frac{14735}{53} = 278;$$

$$F_{0,05}(5; 18) = 6,61; \quad F_{0,01}(5; 18) = 16,26.$$

На любом уровне значимости  $F_{факт.} > F_{табл.}$ , и можно сделать вывод о значимости уравнения регрессии. Статистическая связь между  $y$  и  $x$  доказана.

Величина F- критерия связана с коэффициентом детерминации.

$$F = \frac{r^2}{1-r^2} \cdot (n-2), \quad (14)$$

В линейной регрессии обычно оценивается значимость не только уравнения в целом, но и отдельных его параметров.

Стандартная ошибка коэффициента регрессии определяется по формуле:

$$m_b = \sqrt{\frac{\sum (y - \hat{y}_x)^2 / (n-2)}{\sum (x - \bar{x})^2}} = \sqrt{\frac{S^2}{\sum (x - \bar{x})^2}}, \quad (15)$$

$S^2$  - остаточная дисперсия на одну степень свободы (то же, что и  $D_{остат}$ ).

В рассмотренном примере

$$m_b = \sqrt{\frac{53}{10,857}} = 2,21$$

Величина стандартной ошибки совместно с  $t$ - распределением Стьюдента при  $n-2$  степенях свободы применяется для проверки существенности коэффициента регрессии и для расчета его доверительных интервалов.

Величина коэффициента регрессии сравнивается с его стандартной ошибкой; определяется фактическое значение  $t$ - критерия Стьюдента

$$t_b = \frac{b}{m_b}, \quad (16)$$

которое затем сравнивается с табличным значением при определенном уровне значимости  $\alpha$  и числе степеней свободы ( $n-2$ ). Здесь проверяется нулевая гипотеза в виде  $H_0: b=0$ , также предполагающая несущественность статистической связи между  $y$  и  $x$ , но только учитывающая значение  $b$ , а не соотношение между факторной и остаточной дисперсиями в общем балансе дисперсии результативного признака. Однако общий смысл гипотез один и тот же: проверка наличия статистической связи между  $y$  и  $x$  или её отсутствия.

Если  $t_b > t_{табл}(\alpha; n-2)$ , то гипотеза  $H_0: b=0$  должна быть отклонена, а статистическая связь  $y$  с  $x$  считается установленной. В случае  $t_b < t_{табл}(\alpha; n-2)$  нулевая гипотеза не может быть отклонена, и влияние  $x$  на  $y$  признается несущественным.

В рассмотренном примере

$$t_b = \frac{36,84}{2,21} = 16,67$$

Для двустороннего  $\alpha=0,05$  и  $n-2=5$   $t_{табл}=2,57$ ,  $t_b > t_{табл}$ , поэтому гипотезу о несущественности  $b$  следует отклонить.

Существует связь между  $t_b$  и  $F$ :

$$\begin{aligned} t_b^2 &= \frac{b^2}{m_b^2} = \frac{b^2}{\frac{\sum (y - \hat{y}_x)^2 / (n-2)}{\sum (x - \bar{x})^2}} = \frac{b^2 \sum (x - \bar{x})^2}{\sum (y - \hat{y}_x)^2 / (n-2)} = \\ &= \frac{\sum (x - \bar{x})^2 / 1}{\sum (y - \hat{y}_x)^2 / (n-2)} = \frac{D_{факт.}}{D_{ост.}} = F. \end{aligned}$$

Отсюда следует, что  $t_b^2 = F$ .

Доверительный интервал для  $b$  определяется как  $b \pm t_{табл} \cdot m_b$ . 95%-ные границы в примере составят:

$$36,84 \pm 2,57 \cdot 2,21 = 36,84 \pm 5,68,$$

т.е.  $31,16 \leq b \leq 42,52$ . Это означает, что с вероятностью 0,95 истинное значение  $b$  находится в указанном интервале.

Коэффициент регрессии имеет четкую экономическую интерпретацию, поэтому доверительные границы интервала не должны содержать противоречивых результатов, например,  $-10 \leq b \leq 40$ . Они не должны включать нуль.

Стандартная ошибка параметра  $a$  определяется по формуле:

$$m_a = \sqrt{\frac{\sum (y - \hat{y}_x)^2}{n-2} \cdot \frac{\sum x^2}{n \sum (x - \bar{x})^2}} = \sqrt{S^2 \cdot \frac{\sum x^2}{n \sum (x - \bar{x})^2}} \quad (17)$$

Процедура оценивания существенности  $a$  не отличается от таковой для параметра  $b$ . При этом фактическое значение  $t$ -критерия вычисляется по формуле:

$$t_a = \frac{a}{m_a} \quad (18)$$

Процедура проверки значимости линейного коэффициента корреляции отличается от процедур, приведенных выше. Это объясняется тем, что  $r$  как случайная величина распределена по нормальному закону лишь при большом числе наблюдений и малых значениях  $|r|$ . В этом случае гипотеза об отсутствии корреляционной связи между  $y$  и  $x$   $H_0: r=0$  проверяется на основе статистики

$$t_r = \frac{r}{\sqrt{1-r^2}} \sqrt{n-2}, \quad (19)$$

которая при справедливости  $H_0$  приблизительно распределена по закону Стьюдента с  $(n-2)$  степенями свободы. Если  $t_r > t_{табл}(\alpha; n-2)$ , то гипотеза  $H_0$  отвергается с вероятностью ошибиться, не превышающей  $\alpha$ . Из (19) видно, что в парной линейной регрессии  $t_r^2 = F$ . Кроме того,  $t_b^2 = F$ , поэтому  $t_r^2 = t_b^2$ . Таким образом, проверка гипотез о значимости коэффициентов регрессии и корреляции равносильна проверке гипотезы о существенности линейного уравнения регрессии.

Однако при малых выборках и значениях  $r$ , близких к  $\pm 1$ , следует учитывать, что распределение  $r$  как случайной величины отличается от нормального, и построение доверительных интервалов для  $r$  не может быть выполнено стандартным способом. В этом случае вообще легко прийти к противоречию, заключающемуся в том, что доверительный интервал будет содержать значения, превышающие единицу.

Чтобы обойти это затруднение, используется так называемое z-преобразование Фишера:

$$z = \frac{1}{2} \ln \frac{1+r}{1-r}, \quad (20)$$

которое дает нормально распределенную величину  $z$ , значения которой при изменении  $r$  от  $-1$  до  $+1$  изменяются от  $-\infty$  до  $+\infty$ . Стандартная ошибка этой величины равна:

$$m_z = \frac{1}{\sqrt{n-3}} \quad (21)$$

Для величины  $z$  имеются таблицы, в которых приведены её значения для соответствующих значений  $r$ .

Для  $z$  выдвигается нуль-гипотеза  $H_0: z=0$ , состоящая в том, что корреляция отсутствует. В этом случае значения статистики

$$t_z = \frac{z}{m_z}, \quad (22)$$

которая распределена по закону Стьюдента с  $(n-2)$  степенями свободы, не превышает табличного на соответствующем уровне значимости.

Для каждого значения  $z$  можно вычислить критические значения  $r$ . Таблицы критических значений  $r$  разработаны для уровней значимости 0,05 и 0,01 и соответствующего числа степеней свободы. Если вычисленное значение  $r$  превышает по абсолютной величине табличное, то данное значение  $r$  считается существенным. В противном случае фактическое значение несущественно.

### Интервалы прогноза по линейному уравнению регрессии.

Прогнозирование по уравнению регрессии представляет собой подстановку в уравнение регрессии соответственного значения  $x$ . Такой прогноз  $\hat{y}_x$  называется **точечным**. Он не является точным, поэтому дополняется расчетом стандартной ошибки  $\hat{y}_x$ ; получается **интервальная оценка** прогнозного значения  $y^*$ :

$$\hat{y}_x - m_{\hat{y}_x} \leq y^* \leq \hat{y}_x + m_{\hat{y}_x}$$

Преобразуем уравнение регрессии:

$$\hat{y}_x = a + bx = \bar{y} - b\bar{x} + bx = \bar{y} + b(x - \bar{x})$$

ошибка  $m_{\hat{y}_x}$  зависит от ошибки  $\bar{y}$  и ошибки коэффициента регрессии  $b$ , т.е.  $m_{\hat{y}_x}^2 = m_{\bar{y}}^2 + m_b^2(x - \bar{x})^2$ .

Из теории выборки известно, что  $m_{\bar{y}}^2 = \frac{\sigma^2}{n}$

Используем в качестве оценки  $\sigma^2$  остаточную дисперсию на одну степень свободы  $S^2$ , получаем:  $m_{\bar{y}}^2 = \frac{S^2}{n}$

Ошибка коэффициента регрессии из формулы (15):

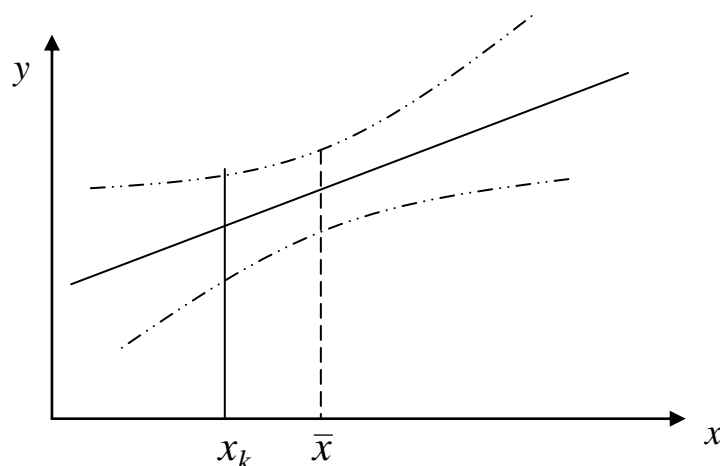
$$m_b^2 = \frac{S^2}{\sum (x - \bar{x})^2}$$

Таким образом, при  $x = x_k$  получаем:

$$m_{\hat{y}_x}^2 = \frac{S^2}{n} + \frac{S^2}{\sum (x - \bar{x})^2} (x_k - \bar{x})^2 = S^2 \left( \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum (x - \bar{x})^2} \right)$$

$$m_{\hat{y}_x} = S \sqrt{\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum (x - \bar{x})^2}} \quad (23)$$

Как видно из формулы (23), величина  $m_{\hat{y}_x}$  достигает минимума при  $x_k = \bar{x}$  и возрастает по мере удаления  $x_k$  от  $\bar{x}$  в любом направлении.



Для нашего примера эта величина составит:

$$m_{\hat{y}_x} = \sqrt{53 \left( \frac{1}{7} + \frac{(x_k - 3,143)^2}{10,857} \right)}$$

При  $x_k = \bar{x}$   $m_{\hat{y}_x} = \sqrt{53 : 7} = 2,75$ . При  $x_k = 4$

$$m_{\hat{y}_x} = \sqrt{53 \left( \frac{1}{7} + \frac{(4 - 3,143)^2}{10,857} \right)} = 3,34$$

Для прогнозируемого значения  $\hat{y}_x$  95% - ные доверительные интервалы при заданном  $x_k$  определены выражением:

$$\hat{y}_{x_k} \pm t_{\alpha} \cdot m_{\hat{y}_x} \quad (24)$$

т.е. при  $x_k = 4$   $\hat{y}_{x_k} \pm 2,57 \cdot 3,34$  или  $\hat{y}_{x_k} \pm 8,58$ . При  $x_k = 4$  прогнозное значение составит  $y_p = -5,79 + 36,84 \cdot 4 = 141,57$  - это точечный прогноз.

Прогноз линии регрессии лежит в интервале:

$$132,99 \leq \hat{y}_{x_k} \leq 150,15$$

Мы рассмотрели доверительные интервалы для **среднего значения  $y$  при заданном  $x$** . Однако фактические значения  $y$  варьиру-

ются около среднего значения  $\hat{y}_x$ , они могут отклоняться на величину случайной ошибки  $\varepsilon$ , дисперсия которой оценивается как остаточная дисперсия на одну степень свободы  $S^2$ . Поэтому ошибка прогноза отдельного значения  $y$  должна включать не только стандартную ошибку  $m_{\hat{y}_x}$ , но и случайную ошибку  $S$ . Таким образом, средняя ошибка прогноза индивидуального значения  $y$  составит:

$$m_{y_{i(x_k)}} = S \sqrt{1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum (x - \bar{x})^2}} \quad (25)$$

Для примера:

$$m_{y_{i(x_k=4)}} = \sqrt{53 \cdot \left( 1 + \frac{1}{7} + \frac{(4 - 3,143)^2}{10,857} \right)} = 8,01$$

Доверительный интервал прогноза индивидуальных значений  $y$  при  $x_k = 4$  с вероятностью 0,95 составит:  $141,57 \pm 2,57 \cdot 8,01$ , или  $120,98 \leq y_p \leq 162,16$ .

Пусть в примере с функцией издержек выдвигается предположение, что в предстоящем году в связи со стабилизацией экономики затраты на производство 8 тыс. ед. продукции не превысят 250 млн. руб. Означает ли это изменение найденной закономерности или затраты соответствуют регрессионной модели?

Точечный прогноз:  $\hat{y}_{x=8} = -5,79 + 36,84 \cdot 8 = 288,93$ .

Предполагаемое значение - 250. Средняя ошибка прогнозного индивидуального значения:

$$\begin{aligned} m_{y_{i(x_i)}} &= S \sqrt{1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum (x - \bar{x})^2}} = \\ &= \sqrt{53 \cdot \left( 1 + \frac{1}{7} + \frac{(8 - 3,143)^2}{10,857} \right)} = 13,26 \end{aligned}$$

Сравним ее с предполагаемым снижением издержек производства, т.е.  $250 - 288,93 = -38,93$ :

$$t = \frac{-38,93}{13,26} = -2,93.$$

Поскольку оценивается только значимость уменьшения затрат, то используется односторонний  $t$ - критерий Стьюдента. При ошибке

в 5 % с  $n-2=5$   $t_{таб.} = 2,015$ , поэтому предполагаемое уменьшение затрат значительно отличается от прогнозируемого значения при 95 % - ном уровне доверия. Однако, если увеличить вероятность до 99%, при ошибке 1 % фактическое значение  $t$  – критерия оказывается ниже табличного 3,365, и различие в затратах статистически не значимо, т.е. затраты соответствуют предложенной регрессионной модели.

### Нелинейная регрессия

До сих пор мы рассматривали лишь **линейную** модель регрессионной зависимости  $y$  от  $x$  (3). В то же время многие важные связи в экономике являются **нелинейными**. Примерами такого рода регрессионных моделей являются производственные функции (зависимости между объемом произведенной продукции и основными факторами производства – трудом, капиталом и т.п.) и функции спроса (зависимости между спросом на какой-либо вид товаров или услуг, с одной стороны, и доходом и ценами на этот и другие товары – с другой).

При анализе нелинейных регрессионных зависимостей наиболее важным вопросом применения классического МНК является способ их линеаризации. В случае линеаризации нелинейной зависимости получаем линейное регрессионное уравнение типа (3), параметры которого оцениваются обычным МНК, после чего можно записать исходное нелинейное соотношение.

Несколько особняком в этом смысле стоит полиномиальная модель произвольной степени:

$$y = a_0 + a_1x + a_2x^2 + \dots + a_kx^k + \varepsilon, \quad (26)$$

к которой обычный МНК можно применять без всякой предварительной линеаризации.

Рассмотрим указанную процедуру применительно к параболе второй степени:

$$y = a + bx + cx^2 + \varepsilon \quad (27)$$

Такая зависимость целесообразна в случае, если для некоторого интервала значений фактора возрастающая зависимость меняется на убывающую или наоборот. В этом случае можно определить значение фактора, при котором достигается максимальное или минимальное значение результативного признака. Если исходные данные не обнаруживают изменение направленности связи, параметры параболы становятся трудно интерпретируемыми, и форму связи лучше заменить другими нелинейными моделями.

Применение МНК для оценки параметров параболы второй степени сводится к дифференцированию суммы квадратов остатков рег-



регрессии по каждому из оцениваемых параметров и приравнению полученных выражений нулю. Получается система нормальных уравнений, число которых равно числу оцениваемых параметров, т.е. трем:

$$\begin{cases} a \cdot n & + b \sum x & + c \sum x^2 & = \sum y, \\ a \sum x & + b \sum x^2 & + c \sum x^3 & = \sum yx, \\ a \sum x^2 & + b \sum x^3 & + c \sum x^4 & = \sum yx^2. \end{cases} \quad (28)$$

Решать эту систему можно любым способом, в частности, методом определителей.

Экстремальное значение функции наблюдается при значении фактора, равном:

$$x = -\frac{b}{2a}.$$

Если  $b > 0$ ,  $c < 0$ , имеет место максимум, т.е. зависимость сначала растет, а затем падает. Такого рода зависимости наблюдаются в экономике труда при изучении заработной платы работников физического труда, когда в роли фактора выступает возраст. При  $b < 0$ ,  $c > 0$  парабола имеет минимум, что обычно проявляется в удельных затратах на производство в зависимости от объема выпускаемой продукции.

В нелинейных зависимостях, не являющихся классическими полиномами, обязательно проводится предварительная линеаризация, которая заключается в преобразовании или переменных, или параметров модели, или в комбинации этих преобразований. Рассмотрим некоторые классы таких зависимостей.

Зависимости гиперболического типа имеют вид:

$$y = a + \frac{b}{x} + \varepsilon \quad (29)$$

Примером такой зависимости является кривая Филлипса, констатирующая обратную зависимость процента прироста заработной платы от уровня безработицы. В этом случае значение параметра  $b$  будет больше нуля. Другим примером зависимости (29) являются кривые Энгеля, формулирующие следующую закономерность: с ростом дохода доля доходов, расходуемых на продовольствие, уменьшается, а доля доходов, расходуемых на непродовольственные товары, будет возрастать. В этом случае  $b < 0$ , а результативный признак в (29) показывает долю расходов на непродовольственные товары.

Линеаризация уравнения (29) сводится к замене фактора  $z = 1/x$ , и уравнение регрессии имеет вид (3), в котором вместо фактора  $x$  используем фактор  $z$ :

$$\hat{y} = a + b \cdot z \quad (30)$$

К такому же линейному уравнению сводится полулогарифмическая кривая:

$$y = a + b \cdot \ln x + \varepsilon \quad (31)$$

которая может быть использована для описания кривых Энгеля. Здесь  $\ln(x)$  заменяется на  $z$ , и получается уравнение (30).

Достаточно широкий класс экономических показателей характеризуется приблизительно постоянным темпом относительного прироста во времени. Этому соответствуют зависимости показательного (экспоненциального) типа, которые записываются в виде:

$$y = e^{a+bx} \cdot \varepsilon \quad (32)$$

или в виде

$$y = a \cdot e^{bx} \cdot \varepsilon \quad (33)$$

Возможна и такая зависимость:

$$y = a \cdot b^x \cdot \varepsilon \quad (34)$$

В регрессиях типа (32) – (34) применяется один и тот же способ линеаризации – логарифмирование. Уравнение (32) приводится к виду:

$$\ln y = a + bx + \ln \varepsilon \quad (35)$$

Замена переменной  $Y = \ln y$  сводит его к линейному виду:

$$Y = a + bx + E, \quad (36)$$

где  $E = \ln \varepsilon$ . Если  $E$  удовлетворяет условиям Гаусса-Маркова, параметры уравнения (32) оцениваются по МНК из уравнения (36). Уравнение (33) приводится к виду:

$$y = \ln a + bx + \ln \varepsilon, \quad (37)$$

который отличается от (35) только видом свободного члена, и линейное уравнение выглядит так:

$$Y = A + bx + E, \quad (38)$$

где  $A = \ln a$ . Параметры  $A$  и  $b$  получаются обычным МНК, затем параметр  $a$  в зависимости (33) получается как антилогарифм  $A$ . При логарифмировании (34) получаем линейную зависимость:

$$Y = a + Bx + E, \quad (39)$$

где  $B = \ln b$ , а остальные обозначения те же, что и выше. Здесь также применяется МНК к преобразованным данным, а параметр  $b$  для (34) получается как антилогарифм коэффициента  $B$ .

Широко распространены в практике социально-экономических исследований степенные зависимости. Они используются для построения и анализа производственных функций. В функциях вида:

$$y = a \cdot x^b \cdot \varepsilon \quad (40)$$

особенно ценным является то обстоятельство, что параметр  $b$  равен коэффициенту эластичности результативного признака по фактору  $x$ . Преобразуя (40) путем логарифмирования, получаем линейную регрессию:

$$Y = A + bX + E \quad (41)$$

где  $Y = \ln y$ ,  $A = \ln a$ ,  $X = \ln x$ ,  $E = \ln \varepsilon$ .

Еще одним видом нелинейности, приводимым к линейному виду, является обратная зависимость:

$$y = \frac{1}{a + bx + \varepsilon} \quad (42)$$

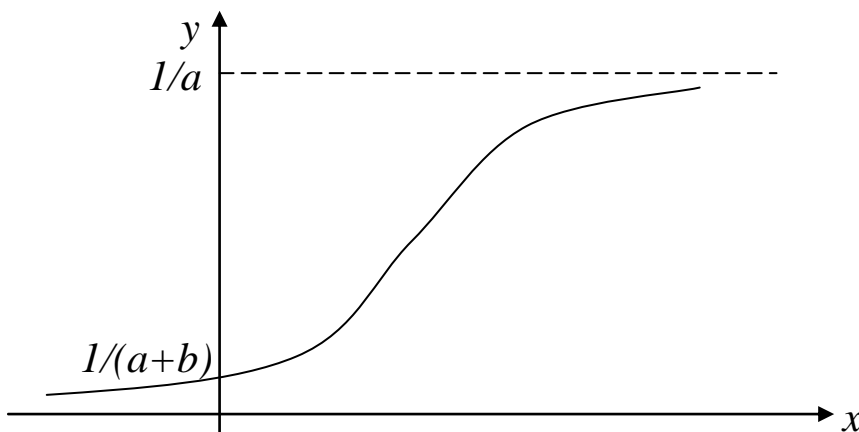
Проводя замену  $u = 1/y$ , получим:

$$u = a + bx + \varepsilon \quad (43)$$

Наконец, следует отметить зависимость логистического типа:

$$y = \frac{1}{a + be^{-x} + \varepsilon} \quad (44)$$

Графиком функции (44) является так называемая «кривая насыщения», которая имеет две горизонтальные асимптоты  $y=0$  и  $y=1/a$  и точку перегиба  $x = \ln(-b/a)$ ,  $y = 1/(-2a)$ , а также точку пересечения с осью ординат  $y = 1/(a+b)$ :



Уравнение (44) приводится к линейному виду заменами переменных  $u = 1/y$ ,  $z = e^{-x}$ .

Любое уравнение нелинейной регрессии, как и линейной зависимости, дополняется показателем корреляции, который в данном случае называется индексом корреляции:

$$R = \sqrt{1 - \frac{\sigma_{ост}^2}{\sigma_y^2}} \quad (45)$$

Здесь  $\sigma_y^2$  - общая дисперсия результативного признака  $y$ ,  $\sigma_{ост}^2$  - остаточная дисперсия, определяемая по уравнению нелинейной регрессии  $\hat{y}_x = f(\hat{x})$ . Следует обратить внимание на то, что разности в соответствующих суммах  $\sum (y - \bar{y})^2$  и  $\sum (y - \hat{y}_x)^2$  берутся не в преобразованных, а в исходных значениях результативного признака. Иначе говоря, при вычислении этих сумм следует использовать не преобразованные (линеаризованные) зависимости, а именно исходные нелинейные уравнения регрессии. По-другому (45) можно записать так:

$$R = \sqrt{1 - \frac{\sum (y - \hat{y}_x)^2}{\sum (y - \bar{y})^2}} \quad (46)$$

Величина  $R$  находится в границах  $0 \leq R \leq 1$ , и чем ближе она к единице, тем теснее связь рассматриваемых признаков, тем более надежно найденное уравнение регрессии. При этом индекс корреляции совпадает с линейным коэффициентом корреляции в случае, когда преобразование переменных с целью линеаризации уравнения регрессии не проводится с величинами результативного признака. Так обстоит дело с полулогарифмической и полиномиальной регрессий, а также с равносторонней гиперболой (29). Определив линейный коэффициент корреляции для линеаризованных уравнений, например, в пакете Excel с помощью функции ЛИНЕЙН, можно использовать его и для нелинейной зависимости.

Иначе обстоит дело в случае, когда преобразование проводится также с величиной  $x$ , например, взятие обратной величины или логарифмирование. Тогда значение  $R$ , вычисленное той же функцией ЛИНЕЙН, будет относиться к линеаризованному уравнению регрессии, а не к исходному нелинейному уравнению, и величины разностей под суммами в (46) будут относиться к преобразованным величинам, а не к исходным, что не одно и то же. При этом, как было сказано выше, для расчета  $R$  следует воспользоваться выражением (46), вычисленным по исходному нелинейному уравнению.

Поскольку в расчете индекса корреляции используется соотношение факторной и общей СКО, то  $R^2$  имеет тот же смысл, что и коэффициент детерминации. В специальных исследованиях величину  $R^2$  для нелинейных связей называют индексом детерминации.

Оценка существенности индекса корреляции проводится так же, как и оценка надежности коэффициента корреляции.

Индекс детерминации используется для проверки существенности в целом уравнения нелинейной регрессии по  $F$ -критерию Фишера:

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m}, \quad (47)$$

где  $n$ -число наблюдений,  $m$ -число параметров при переменных  $x$ . Во всех рассмотренных нами случаях, кроме полиномиальной регрессии,  $m=1$ , для полиномов (26)  $m=k$ , т.е. степени полинома. Величина  $m$  характеризует число степеней свободы для факторной СКО, а  $(n-m-1)$  – число степеней свободы для остаточной СКО.

Индекс детерминации  $R^2$  можно сравнивать с коэффициентом детерминации  $r^2$  для обоснования возможности применения линейной функции. Чем больше кривизна линии регрессии, тем больше разница между  $R^2$  и  $r^2$ . Близость этих показателей означает, что усложнять форму уравнения регрессии не следует и можно использовать линейную функцию. Практически, если величина  $(R^2 - r^2)$  не превышает 0,1, то линейная зависимость считается оправданной. В противном случае проводится оценка существенности различия показателей детерминации, вычисленных по одним и тем же данным, через  $t$ -критерий Стьюдента:

$$t = \frac{R^2 - r^2}{m_{|R-r|}} \quad (48)$$

Здесь в знаменателе находится ошибка разности  $(R^2 - r^2)$ , определяемая по формуле:

$$m_{|R-r|} = 2 \cdot \sqrt{\frac{\left( (R^2 - r^2) - \left( (R^2 - r^2) \right)^2 \cdot \left( 2 - (R^2 + r^2) \right) \right)}{n}} \quad (49)$$

Если  $t > t_{\text{табл}}(\alpha; n - m - 1)$ , то различия между показателями корреляции существенны и замена нелинейной регрессии линейной нецелесообразна.

В заключение приведем формулы расчета коэффициентов эластичности для наиболее распространенных уравнений регрессии:

Вид уравнения регрессии	Коэффициент эластичности
$y = a + b \cdot x + \varepsilon$	$\frac{b \cdot x}{a + bx}$
$y = a + bx + cx^2 + \varepsilon$	$\frac{b + 2cx}{a + bx + cx^2}$

$y = a + \frac{b}{x} + \varepsilon$	$\frac{-b}{ax + b}$
$y = a \cdot b^x \cdot \varepsilon$	$x \ln b$
$y = a \cdot x^b$	$b$
$y = a + b \ln x + \varepsilon$	$\frac{b}{a + b \ln x}$
$y = \frac{1}{a + bx + \varepsilon}$	$\frac{-bx}{a + bx}$

### Список учебной литературы

1. Эконометрика: Учебник /Под ред. И.И. Елисеевой/ - М.: Финансы и статистика, 2001. – 344с.
2. Практикум по эконометрике: Учебное пособие / И.И. Елисеева и др./ - М.: Финансы и статистика, 2001. – 192с.
3. Бородич С.А. Эконометрика: Учебное пособие. – М.: Новое знание. 2001. – 408с.
4. Магнус Я.Р., Катышев П.К., Пересецкий А.А., Эконометрика. Начальный курс. Учебное пособие. – М.: Дело, 1998. – 248с.
5. Доугерти К. Введение в эконометрику. – М.: ИНФРА-М, 1997. – 402с.